DOI: https://doi.org/10.52756/ijerr.2024.v46.003 *Int. J. Exp. Res. Rev.*, Vol. 46: 31-44(2024)

Original Article | Peer Reviewed | Open Access

International Journal of Experimental Research and Review (IJERR)
© Copyright by International Academic Publishing House (IAPH)
ISSN: 2455-4855 (Online)
www.iaph.in

Analysis and Visualize the Predictive Model Performance: Manual Vs Automated Machine Learning (AutoML) Algorithms for Heart Failure Prediction

C. Rajeev[1]* and Karthika Natarajan[2]

[1]Department of School of Computer Science and Engineering, VIT-AP University, Amaravati-522237, Andhra Pradesh, India; [2]Department of School of Computer Science and Engineering, VIT-AP University, Amaravati-522237, Andhra Pradesh, India

E-mail/Orcid Id:

CR, rajeev.21phd7135@vitap.ac.in, https://orcid.org/0000-0002-9942-6452; *KN*, Karthika.n@vitap.ac.in, https://orcid.org/0000-0002-6700-583X

Article History:

Received: 20th Apr., 2024
Accepted: 07th Dec., 2024
Published: 30th Dec., 2024

Keywords:

AutoML, AutoGluon, Heart Failure, Machine Learning, SHapley Additive exPlanations

How to cite this Article:
C. Rajeev and Karthika Natarajan (2024). Analysis and Visualize the Predictive Model Performance: Manual Vs Automated Machine Learning (AutoML) Algorithms for Heart Failure Prediction. *International Journal of Experimental Research and Review*, *46*, 31-44.
DOI: https://doi.org/10.52756/ijerr.2024.v46.003

Abstract: Heart failure (HF) is a common complication of cardiovascular diseases. This research focuses on assessing the effectiveness of different models for predicting HF using both Traditional Machine Learning (TML) methods and Automated Machine Learning (AutoML) approaches. TML models need extensive manual tuning and expert knowledge for algorithm selection and optimization, making the process slow and susceptible to human error. To tackle this challenge, the work proposed an AutoML approach utilizing the AutoGluon framework for predicting HF. The main goal of this study is to automate the process of selecting the most efficient model. This study compares a total of twenty (20) individual-trained ML models, consisting of fourteen (14) from AutoML and six (6) from TML. In TML, Logistic Regression (LR) produced the highest 87.50% accuracy and ROC-AUC of 88.83% compared to Support Vector Models (SVM), Decision Trees (DT), Gaussian Naïve Bayes (GNB), Random Forests (RF) and K-Nearest Neighbors (KNN). In AutoML, the CatBoost model outperforms the other thirteen algorithms with the highest accuracy of 99.39% and ROC-AUC of 99.89%. The results show that an AutoML based algorithm called the CatBoost model gives the most accurate model among all 20 models. SHAP was employed to interpret the top-performing model, increasing its transparency and usability.

Introduction

The heart is a crucial organ necessary for the proper functioning of the body, as it pumps blood throughout the system, which is vital for sustaining health (David et al., 2018). The fast-paced nature of contemporary life has resulted in an increase in heart-related health problems. As a result, heart disease has emerged as a significant global health issue, greatly impacting illness and death rates. This highlights the important role of the healthcare system in addressing these serious health challenges. This study focuses on HF, a condition where the heart weakens and can't pump blood effectively enough for the body's needs. HF is not a single condition but rather a category with different types based on various factors. Though symptoms may overlap, accurately identifying the specific type is crucial for doctors to choose the most effective treatment plan. HF arises from various culprits (Hajouli et al., 2022). Underlying conditions like high BP, diabetes, lung problems (chronic respiratory disease), heart muscle disease (cardiomyopathy), and clogged arteries (coronary artery disease) can all contribute. Risk factors (National Heart, Lung and Blood Institute 2018) include a family history of heart problems, age (especially seniors), African-American ethnicity, and unhealthy habits like excessive alcohol consumption, tobacco use, and substance abuse. HF is divided into two primary categories based on how it affects heart pumping. Systolic HF, also known as HF with a low ejection fraction, occurs when the heart muscle weakens and struggles to squeeze. When the heart muscle stiffens and strains to relax and fill with blood while resting, diastolic HF ensues.

Healthcare professionals are essential in diagnosing HF through comprehensive evaluations, utilizing

*Corresponding Author: rajeev.21phd7135@vitap.ac.in advanced imaging techniques, and performing diagnostic procedures, including ECG, brain natriuretic peptide test, chest X-ray, etc. However, these procedures (Shah et al., 2020) take a lot of time, money, and effort from both individuals and healthcare professionals, and they do not always find the exact type of heart disease at the early stages. The rise of HF in young people, the financial strains, the shortcomings of current medical equipment, and the difficulties in diagnosis emphasize the need for creative solutions. Computerized techniques have evolved as viable alternatives to conventional approaches, providing more rapid and efficient HF risk prediction. Advanced computer-aided technologies, including Machine Learning (ML), Deep Learning (DL), and AutoML, hold significant promise in improving the early detection and diagnosis of HF. These techniques offer a more effective way to address the challenges posed by this complex condition compared to traditional methods.

Here is a summary of the study's main contribution:

1) Several predictive models were built using both TML methods, including SVM, LR, DT, KNN, RF, GNB, and the AutoML framework AutoGluon.
2) The research utilizes a dataset of heart information consisting of 303 patient records acquired from Kaggle.
3) The study selects, evaluates, and validates the best-performing models from both TML and AutoML approaches, comparing their prediction results using performance metrics such as accuracy, ROC-AUC, and so on.
4) SHAP was used to analyze AutoML models to explain the predictions and focus on the most important predictive variables.

The organization of this research is as follows: Section 2 conducts a literature review, while Section 3 offers an overview of TML models and AutoML, along with a detailed description of the proposed model. Section 4 discusses the results obtained from the implementation of this approach. Finally, Section 5 provides a conclusion and future research.

Background work

People widely recognize HF as a leading cause of mortality. Recent research has focused extensively on using ML to predict HF, aiming to improve patient outcomes and healthcare management. Numerous studies have investigated various techniques, datasets, and performance metrics to enhance the accuracy and reliability of these prediction models. Traditional methods for diagnosing heart disease have typically relied on a patient's medical history, physical examinations, and symptom assessments by medical professionals. Among these methods, angiography is considered highly accurate for identifying heart conditions. However, angiography is associated with drawbacks such as high costs, potential side effects and the requirement for specialized technical skills (Patil et al., 2009). To overcome these issues, numerous researchers used ML models such as SVM, DT, and so on (Detrano et al., 2009). Krittanawong et al. (2019) evaluated and compared their research using ML classifiers such as LR, RF and DT. Their findings indicated that ML algorithms could significantly improve their ability to predict HF, with RF demonstrating the best results. According to most researchers, the three most commonly used methods are DT, ANN, and SVM to predict heart disease.

Again, researchers have investigated the application of ML to heart diagnosis, utilizing various methods such as feature selection, feature analysis, hyperparameter tuning, balancing techniques, hybrid models, and ensemble techniques to achieve improved results. The author (Ranganathan et al., 2024) used statistical analysis such as Pearson correlation analysis to provide the relationship between the features in their dataset and improve accuracy. The author (Mohan et al., 2019) proposed a novel method of hybrid RF with a linear model to increase accuracy and identify significant features in heart disease. Gardner et al. (1984) combined RF, Gradient Boost (GB), and KNN into an ensemble model to improve accuracy and robustness over the individual models. Tarawneh et al. (2019) examined a range of studies on both single-model and hybrid-model approaches, and their findings indicate that hybrid models exhibit superior accuracy in disease prediction compared to single models. Baseer et al. (2023) conducted a comparison with and without hyperparameter tuning using ML algorithms for HF prediction. Using GridSearchCV with SVM results in a significantly higher accuracy of 99.02%, compared to the 74% accuracy achieved by SVM without GridSearchCV.

Khourdifi et al. (2019) enhanced the performance of ML models by integrating particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) to identify important features and improve accuracy. Gazeloglu et al. (2020) compared the prediction of CVD with and without feature selection. In this case, Naïve Bayes (NB) and Fuzzy Rough Sets achieve the highest accuracy compared to those without feature selection when combined with correlation-based Feature Selection (CFS). Shah et al. (2017) employed Principal Component Analysis (PCA)

DOI: https://doi.org/10.52756/ijerr.2024.v46.003

for feature selection. Bodapati et al. (2019) proposed a clustering-based method for extracting pertinent features that significantly influence classification outcomes. They employed both K-means clustering and spectral clustering for cluster formation to mitigate clustering uncertainty. Nasarian et al. (2020) created heterogeneous hybrid feature selection (2HFS) to discover critical CAD features. ADASYN and SMOTE were used to address the dataset imbalance. This method improved classification accuracy to 92.58%. Waqar et al. (2021) introduced a SMOTE-based artificial neural network to address imbalanced data, eliminating the necessity for feature engineering in datasets. This approach surpassed the performance of all other models.

Many researchers have delved into utilizing ML to diagnose HF through various models. However, incorporating ML often demands a level of computer science proficiency, potentially hindering widespread adoption among healthcare practitioners (Ferreira et al., 2021). Standard stages in a ML endeavor encompass defining the problem, acquiring data, conducting exploratory data analysis, data preparation, model exploration, and model refinement. AutoML (Absar et al., 2020) platforms make it easier to use advanced models by automating feature engineering and hyperparameter tuning. Compared to TML methods, these platforms require a lot less code and technical knowledge. These platforms automatically handle data preparation, model selection, and model refinement processes. AutoML frameworks provide healthcare providers with a cost-effective tool for identifying and predicting cardiac diseases.

Paladino et al. (2023) explored the effectiveness of AutoML tools in diagnosing heart disease. They used AutoGluon, PyCaret, and AutoKeras to generate predictive models to achieve this. The results indicated that AutoGluon consistently outperformed the other tools, achieving an accuracy of 86%. Orlenko et al. (2020) used the TPOT tool to predict CAD diagnosis more accurately and in less time. They found that TPOT's automated optimization produced better predictive models than grid search. Auto-Sklearn beat TML on two cardiovascular datasets, according to Padmanabhan et al. (2019) and Pol et al. (2021) used PyCaret to forecast heart disease. Ferreira et al. (2021) conducted a comparative study that evaluated eight AutoML tools (rminer, AutoKeras, AutoSklearn, H2O, TPOT, TransmogrifAI, AutoGluon, and PyTorch) across three scenarios: GML, DL, and XGB. As confirmed by OpenML results, contemporary GML AutoML algorithms outperformed human ML modeling on five datasets. Rimal et al. (2023) conducted a comparison between conventional ML models and AutoML in the context of heart disease classification. Among all models, the AutoML-produced generalized linear model demonstrated the highest accuracy.

The advantage of using AutoML for model building is its ability to efficiently optimize hyperparameters in a short time, with the option to set time constraints on execution duration. AutoML automatically identifies the algorithms that provide the most accurate predictions for the given dataset. ML models can be used to identify important HF predictors and distinguish between individuals with and without an HF diagnosis.

Materials and Methods

Traditional Machine Learning (TML)

TML has seen significant advancements in recent years, as outlined by Tufail et al. (2023), involving several steps crucial for developing ML models, each requiring manual intervention, as depicted in figure 1 (right part). This figure illustrates the basic architecture of both AutoML and TML. In TML, the process begins with collecting high-quality, relevant data from various sources like databases, APIs, or public repositories. Next, the data undergoes manual preprocessing, including cleaning, normalizing, scaling, and encoding. Feature selection follows, using techniques such as correlation, PCA, etc., to identify the important and pertinent features. The data is then manually split into training and testing sets, typically in ratios like 80-20 or 70-30, to ensure the model can generalize well. Choosing the appropriate ML algorithm depends on the specific tasks, such as classification or regression. Finally, the model performance is manually evaluated using various metrics like F1-score, accuracy, precision, Mean Squared Error (MSE), or R-squared recall, often employing cross-validation for robustness. Despite these advancements, developing ML models traditionally remains resource-intensive, requiring significant expertise and time.

AutoML

The complexity of cutting-edge TML techniques is constantly evolving, making it difficult for ML experts to incorporate the latest best practices into their models. To address this challenge, this study utilized AutoML. AutoML (He et al., 2020) streamlines and automates the complete process of applying ML to practical problems, as depicted in figure 1 (left part). By handling complex tasks, AutoML makes ML more accessible to a wider audience, including those with limited expertise. It cuts down on the time and effort needed to create effective models by providing user-friendly interfaces or APIs that allow users to input their data and receive optimized

DOI: https://doi.org/10.52756/ijerr.2024.v46.003

models without requiring extensive knowledge of the underlying algorithms and techniques. AutoML simplifies the ML process by automating the selection of appropriate algorithms (Rajeev et al., 2024) and hyperparameters, and it excels at optimizing these hyperparameters, a crucial and often time-consuming aspect of achieving optimal model performance.

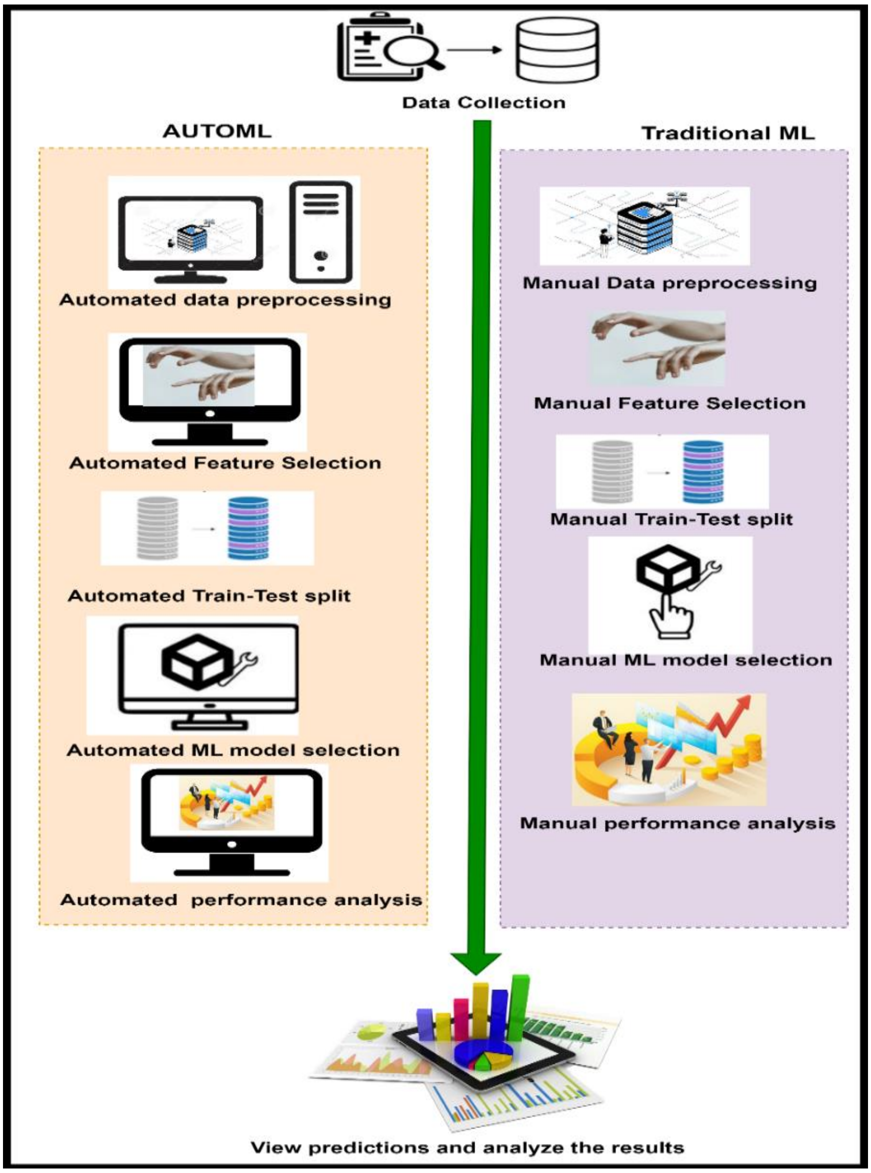

Data Collection
AUTOML
Traditional ML
Automated data preprocessing
Automated Feature Selection
Automated Train-Test split
Automated ML model selection
Automated performance analysis
Manual Data preprocessing
Manual Feature Selection
Manual Train-Test split
Manual ML model selection
Manual performance analysis
View predictions and analyze the results

Figure 1. Basic architecture of AutoML vs TML.

Figure 1 illustrates (left part) the basic architecture of AutoML, highlighting its differences from the TML framework. According to Shen et al. (2018), the general steps of AutoML include data collection from various sources, similar to TML. AutoML then automates data preprocessing, which involves data cleaning, handling missing values, and formatting it for analysis, saving significant time and effort. It also automatically selects features from the data. It also automatically divides the data into training and testing sets for performance evaluation. It also selects the most suitable ML model for the specific task and data, eliminating the need for manual model selection. It also automates performance analysis, typically calculating accuracy, precision and recall metrics. Finally, the trained model can be applied to new data for predictions, enabling researchers to evaluate the outcomes and pinpoint areas that need enhancement. Several AutoML tools are available to automate model development. Leading AutoML platforms include Auto-PyTorch, Microsoft Azure AutoML, H_2O and AutoGluon. In this work, AutoGluon is utilized.

Proposed Methodology

Figure 2 presents the proposed methodology for HF prediction, which includes the following steps: 1) Dataset Overview, 2) Dataset Splitting, 3) AutoGluon, and 4) SHAP Analysis, as described below:

Heart disease dataset

DOI: https://doi.org/10.52756/ijerr.2024.v46.003

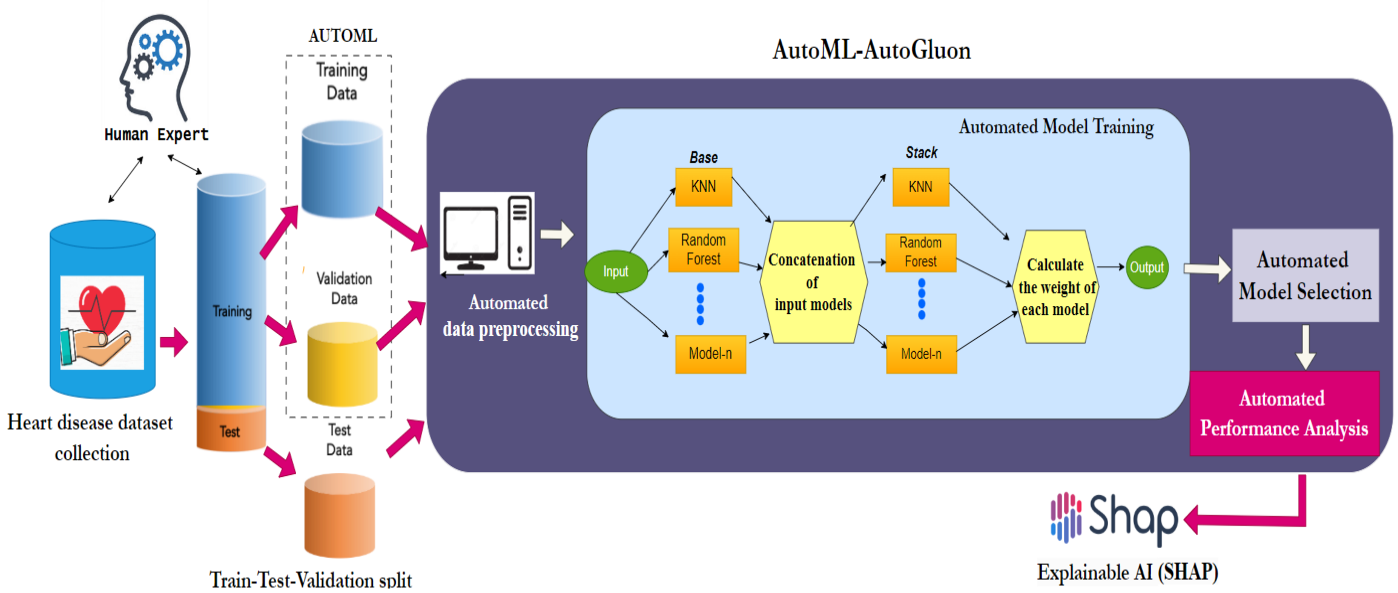

Human Expert
AUTOML
Training Data
Training
Validation Data
Test
Test Data
Heart disease dataset collection
Train-Test-Validation split
AutoML-AutoGluon
Automated data preprocessing
Automated Model Training
Base
Stack
Input
KNN
Random Forest
Model-n
Concatenation of input models
Calculate the weight of each model
Output
Automated Model Selection
Automated Performance Analysis
Shap
Explainable AI (SHAP)

Figure 2. Proposed HF prediction system.

This work uses the "heart disease" dataset from the UCI repository (Janosi et al., 1988). It is composed of 1,025 records and 14 parameters. This dataset, which is free from missing or null values, serves as the basis for the HF prediction experiment, i.e., a clean dataset. Table 1 provides a detailed description of the features utilized in the study.

Table 1. Dataset features and their description.

Feature	Description
age	Age in years
sex	Sex (1 = male; 0 = female)
cp	Chest pain type
trestbps	Resting blood pressure (in mm Hg)
chol	Serum cholesterol in mg/dL
fbs	Fasting blood sugar > 120 mg/dL (1 = true; 0 = false)
restecg	Resting electrocardiographic results
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment
ca	Number of major vessels (0–3) colored by fluoroscopy

Dataset splitting

ML fundamentally divides a dataset into training and test sets to enhance accuracy and prevent overfitting. A human expert split this heart disease dataset of 1025 records into 80% (820 records) training data and 20% (205 records) test data. This test data was set aside for final performance evaluation. After that, AutoML divided the training data into two groups: a training group with 80% (656 records) for model building and a validation group with 20% (164 records) for hyperparameter tuning and overfitting prevention during the ML process, as illustrated in figure 3.

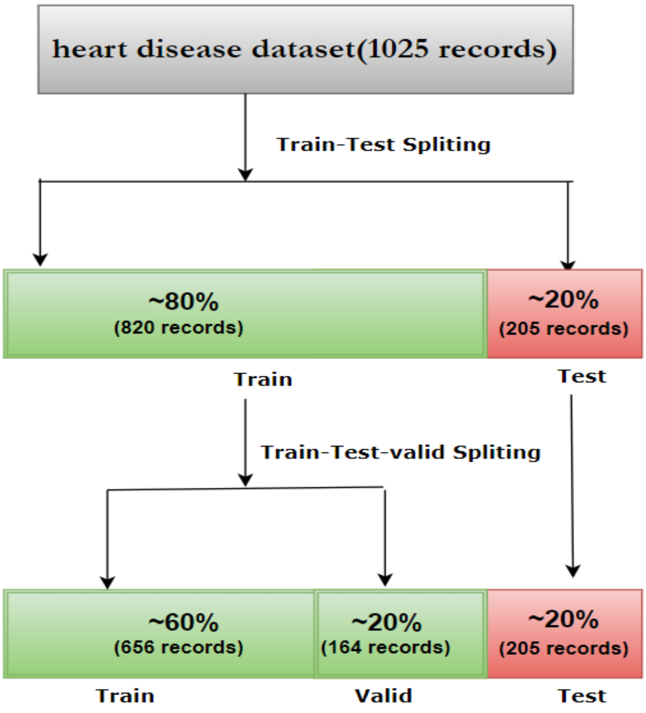


Figure 3. Data splitting by AutoML.

AutoML-AutoGluon

Amazon Web Services (AWS) designed this AutoGluon library to streamline and speed up the development and deployment of ML models. This tool uses text, images, and tabular data. It automates various steps in the ML pipeline described below:

Automated Data preprocessing

AutoGluon's robust automated data preprocessing capabilities simplify and expedite data preparation for ML tasks. It automatically detects and handles missing values in the data with mean, median, or other statistical measures. In this work, AutoGluon ensures that there are

DOI: https://doi.org/10.52756/ijerr.2024.v46.003

no missing values in the data. It also automatically performs categorical encoding using one-hot or ordinal encoding without human intervention. This encoding is used to convert categorical variables into numerical values based on their order or rank. Table 2. provides the parameters of data preprocessing automated by AutoGluon.

Table 2. Automated Data preprocessing parameters.

Parameter	Description
Missing values	No
Train data instances	820
Valid data instances	205
Columns of Train Data	13
Target Column	target
Type of Problem	binary
Target mapping	class 0 = 0, class 1 = 1

Automated model training

The AutoGluon strategy (Erickson et al., 2020) introduces a novel form of multi-layer stack ensembling approach and n-repeated k-fold bagging to improve model performance. Algorithm 1 summarizes this approach. Initially, the training data (A, B) undergoes preprocessing to extract relevant features, employing normalization and feature engineering methods. The family of models (M_f) encompasses the set of algorithms for training, and L indicates the number of stacking layers. Initially, the algorithm follows a stacking loop, iterating through each layer. An n-repeated procedure splits the data into k parts within this loop, forming k folds for cross-validation. Each model type m in M_f is trained on k-1 folds and validated on the remaining fold, producing out-of-fold (OOF) predictions (Sun et al., 2023). These OOF predictions are averaged over all n repetitions and k folds to minimize variance. The averaged OOF predictions are then merged with the original feature matrix, creating an expanded dataset for the next stacking layer. This process is repeated for all specified stacking layers, resulting in a robust ensemble model. In this work, the stacking process employs two layers and uses k=5 subsets for cross-validation.

Algorithm 1 AutoGluon training strategy (Erickson et al., 2020)

```
Impose: Datapoint (A, B); Models family Mf ; Number of layers L
Step 1:  To extract features, do the data preprocess
Step 2: for z= 1 to L do        #Stacking
Step 3:     for x= 1 to n do   #n_repeated
Step 4:         Divide data into k chunks randomly $\{A^y, B^y\}_{y=1}^{k}$
Step 5:         for y= 1 to k do        # k fold bagging
Step 6:             for each model type m in Mf  do
Step 7:                 Train a model type m on A-y, B-y
Step 8:                 Generate predictions $B_{m,x}^{\wedge y}$ on OOF data  Xy
Step 9:             end for
Step 10:        end for
Step 11:    end for
Step 12:    Average OOF predictions  $\hat{B}_m = \{\frac{1}{n}\sum_x B_{m,x}^{\wedge y}\}_{y=1}^{k}$
Step 13:    A ← Concatenate (A, $\{\hat{B}_m\}$  m ∈M )
Step 14: end for
```

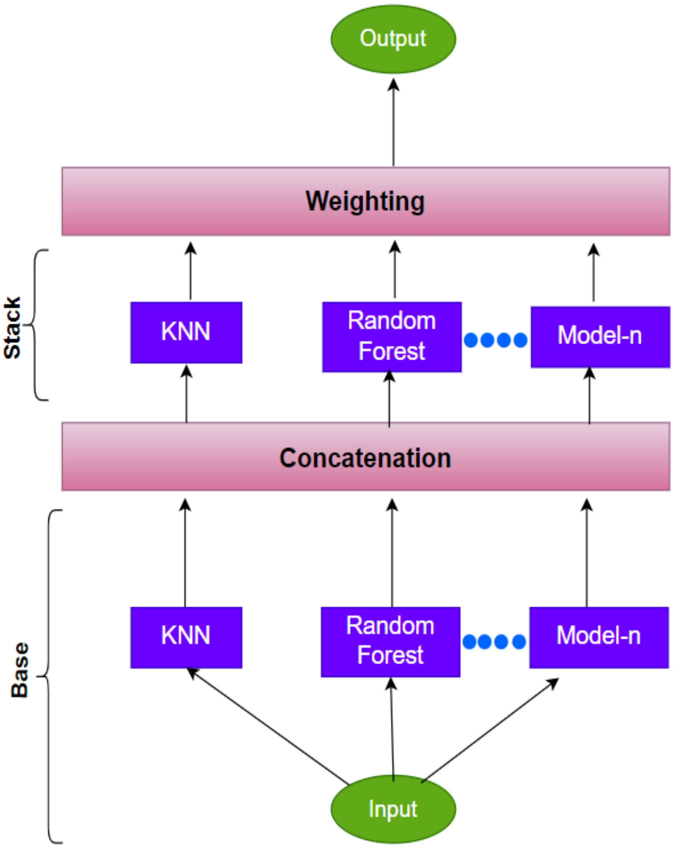


Figure 4. Multi-layer stacking ensemble strategy using AutoGluon.

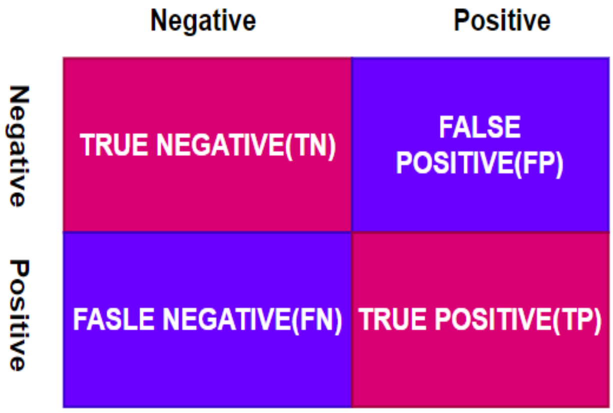
Figure 4 illustrates AutoGluon's multi-layer stacking ensemble strategy, which includes two stacking layers. Initially, the dataset is input into various base models, such as RF, KNN, CatBoost, LightGBM, XGBoost, Extra Trees and NN. The outputs from these base models are concatenated and fed into the next layer, which comprises multiple stacker models. These stacker models function as base models for this layer, learning from the aggregated predictions of the initial base models. The predictions from the stacker models are then combined using a weighting mechanism that assigns different weights to each model's predictions based on their performance, resulting in the final prediction.

Automated Model Selection

AutoGluon automates model selection using a multi-stacking approach. This work configures the tool to generate 14 models by setting the parameter models=14 in the fit() function. Researchers can specify any number of models for experimentation. The models produced include CatBoost, RandomForestGini, RandomForestEntr, NeuralNetTorch, LightGBMLarge, NeuralNetFastAI, ExtraTreesEntr, ExtraTreesGini, LightGBM, WeightedEnsemble_L2, KNeighborsDist, XGBoost, LightGBMXT, and KNeighborsUnif. The leaderboard showcases the performance metrics of all 14 models trained by AutoGluon. Among these algorithms, CatBoost stands out as the top performer, as indicated by the various metrics detailed in Table 2.

Automated performance analysis

AutoGluon automatically offers a variety of evaluation metrics to help assess model performance. AutoGluon tailors the selection of metrics to the specific type of ML task, such as classification or regression. In this work, which focuses on the classification problem, the evaluation utilized accuracy (A), precision (P), Recall (R), F1-score (F1), and the ROC-AUC curve. These metrics provide a comprehensive view of the model's effectiveness in binary classification tasks. Figure 5 showcases the confusion matrix, a performance evaluation, by comparing the model's predictions with the actual class labels.



	Negative	Positive
Negative	TRUE NEGATIVE(TN)	FALSE POSITIVE(FP)
Positive	FASLE NEGATIVE(FN)	TRUE POSITIVE(TP)

Figure 5. Structure of confusion matrix.

A True Positive (TP) is when a person with HF is correctly diagnosed as having HF, whereas a True Negative (TN) is when someone without heart failure is accurately identified as not having HF. In contrast, a False Positive (FP) occurs when a person without heart failure is wrongly diagnosed with HF, and a False Negative (FN) happens when someone with heart failure is incorrectly identified as not having HF.

Accuracy (A) (Pal et al., 2022) measures how accurately it predicts the outcome. Equation (1) expresses the accuracy as the ratio of the count of accurate predictions to the total count of predictions. However, if the data isn't balanced, this statistic might be biased and provide distorted findings.

$$A = \frac{TP+TN}{FP+FN+TP+TN} \qquad (1)$$

According to Equation (2), Precision(P) is defined as the ratio of TP to the total of TP and FP, or the number of TP divided by the total. Equation (3) defines Recall(R) as the ratio of the TP of all True Positives, including FN. The F1 score (F1) is calculated by taking the harmonic mean of the P and R variables, which equals two times the product of the two variables distributed by the total of the two variables (Equation 4).

$$P = \frac{TP}{TP+FP} \qquad (2)$$

$$R = \frac{TP}{TP+FN} \qquad (3)$$

$$F1 = 2 * \frac{P*R}{P+R} \qquad (4)$$

The ROC curve (Deepa et al., 2024) illustrates the relationship between the False Positive Rate (FPR) and the R across various threshold settings to assess the model's accuracy. A higher ROC-AUC score signifies a greater ability to distinguish between classes. An area under the curve (AUC) of 1 indicates a perfect classifier, whereas an area of 0.5 indicates a performance equivalent to random guessing.

Explainable AI (XAI) Integration

XAI (ElShawi et al., 2020) pertains to the advancement of AI systems designed to deliver precise predictions or decisions and to furnish clear and comprehensible rationales for their results. The goal is to enhance trust, accountability, and user acceptance by allowing humans, whether end-users or domain experts, to understand the process and reasoning behind a specific decision or prediction. Local and global interpretability approaches broadly categorize the various XAI methods. Local interpretability focuses on understanding individual predictions and providing explanations for specific instances. Techniques such as LIME and SHAP are frequently used to demonstrate how specific input features impact individual predictions, thus improving the transparency of the model's decision-making on a case-by-case basis. In contrast, global interpretability aims to provide a broad understanding of the model's overall behavior. This approach looks at general patterns and relationships within the model, offering insights into feature importance, model rules, and the overall decision logic. Methods like DT, feature importance scores, and partial dependence plots help illustrate how the model makes predictions across the whole dataset, ensuring a broader comprehension of the model's functionality. This study uses only SHAP for global interpretability.

DOI: https://doi.org/10.52756/ijerr.2024.v46.003

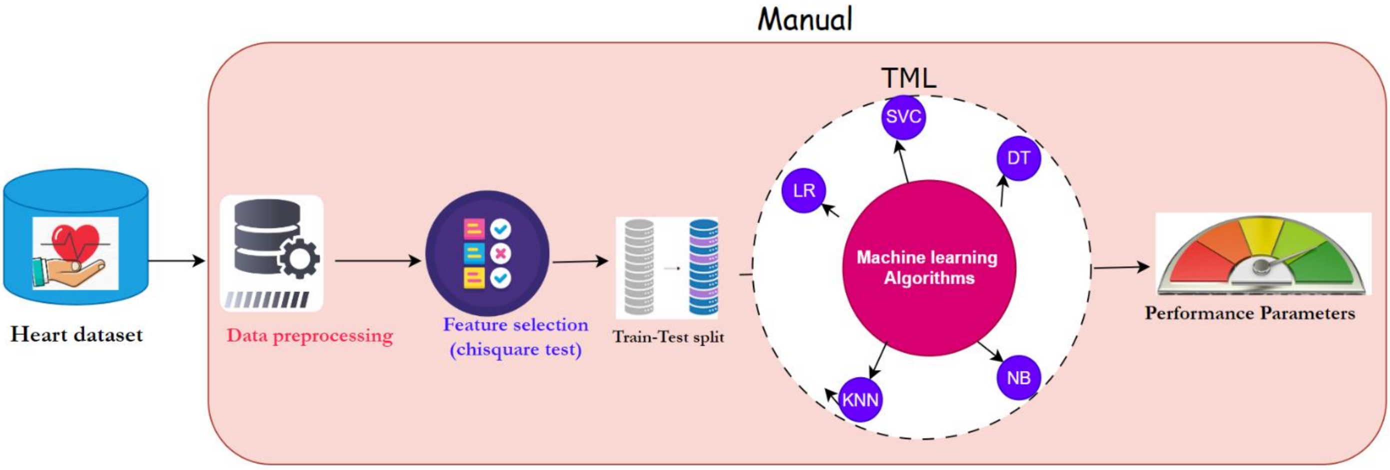


Figure 6. TML for HF prediction.

Shapley Additive ExPlanations (SHAP)

Its goal is to provide insights into the overall behavior of the model across its entire dataset and features. SHAP (Jiang et al., 2023) assigns a value to each feature in a prediction, representing its contribution to the overall model output. Shapley values aim to fairly and interpretably distribute the model's prediction among its features. The results and discussions provide a clear explanation of these predictions through visualization. For global interpretability, SHAP is employed as a tool. SHAP values offer a methodology for interpreting the outcomes of ML models, shedding light on the individual contributions of features to predictions for specific instances. SHAP values associated with individual features can be either positive or negative. Positive SHAP values indicate that a feature contributes to an increase in the prediction, while negative SHAP values suggest that a feature contributes to a decrease in the prediction. By aggregating SHAP values across all instances, a more comprehensive understanding of the importance of features for the classifier is achieved. This approach aids in the interpretation of the classifier's predictions and the identification of key features influencing its behavior. As a result, SHAP values play a critical role in providing deeper insight into the functioning of classifiers, making them useful for interpreting diverse model outputs.

The SHAP value is calculated using equation (5). This tool enables the assessment of the correlation between the model's prediction and the constituent elements employed to attain such a predicted value.

$$\Phi(c,x) = \sum_{z'\subseteq x'} \frac{|z'|!(T-|z'|-1)!}{T!} \left[c_x(z) - c_{\dot{x}}\left(\frac{z'}{i}\right)\right] \quad (5)$$

Where the shapely value for the feature is defined by the function Φi(c, x) of the shape. i and x are vectors that represent the feature value and represent all the possible combinations of the feature subset. This is done to demonstrate the interactions that occur between such individual feature values. The symbol x denotes the simplified data input, and z is the total number of columns in the dataset. SHAP results are plotted and discussed in section 4.

TML for HF prediction

Figure 6. employs several proven steps to predict heart failure using TML. In the TML, data preprocessing for heart disease involved min-max normalization. The results indicated that there were no missing or null values in the data. Feature selection crucial for enhancing model performance was performed using the Chi-square test. This test helps identify features significantly related to the target variable, which is especially useful for categorical data. In this case, the test indicated that all features were critical for predicting HF, except for RestingECG. The data was then split into training and testing sets in an 80-20 ratio. Subsequently, various TML algorithms were applied, including LR, SVC, DT, KNN, and GNB. Among these, LR emerged as the best model, with an accuracy of 87.50% (Natarajan et al., 2024). However, TML is labor-intensive and requires substantial expertise. Data preprocessing involves manual feature engineering and selection, demanding domain knowledge and time. Model selection often relies on trial-and-error approaches for algorithm selection and hyperparameter tuning, which can be time-consuming and prone to human error. Scaling traditional ML to handle large datasets or multiple models is challenging, requiring significant computational resources and manual effort. Reproducing results can be difficult due to inconsistencies in pipelines and documentation. Despite these challenges, TML methods offer control and

Table 3. Comparison of TML vs AutoGluon using various performance parameters.

	Algorithms	Accuracy	Precision	Recall	F1_score	ROC_AUC
AUTOGLUON	***CatBoost***	**99.02%**	**98.15%**	**100.00%**	**99.07%**	**99.89%**
	RandomForestGini	99.02%	98.15%	100.00%	99.07%	99.76%
	RandomForestEntr	99.02%	98.15%	100.00%	99.07%	99.75%
	NeuralNetTorch	98.05%	98.11%	98.11%	98.11%	98.59%
	LightGBMLarge	98.05%	98.11%	98.11%	98.11%	99.31%
	NeuralNetFastAI	97.07%	96.30%	98.11%	97.20%	99.10%
	ExtraTreesEntr	97.07%	98.08%	96.23%	97.14%	99.87%
	ExtraTreesGini	97.07%	98.08%	96.23%	97.14%	99.85%
	LightGBM	95.12%	96.15%	94.34%	95.24%	98.01%
	WeightedEnsemble_L2	95.12%	96.15%	94.34%	95.24%	98.01%
	KNeighborsDist	94.15%	97.96%	90.57%	94.12%	96.86%
	XGBoost	94.15%	93.52%	95.28%	94.39%	97.33%
	LightGBMXT	90.24%	90.57%	90.57%	90.57%	97.64%
	KNeighborsUnif	63.90%	63.33%	71.70%	67.26%	76.79%
TML	***LR***	**87.50%**	**86.73%**	**89.47%**	**88.08%**	**88.43%**
	SVC	87.50%	85.73%	84.47%	87.08%	87.43%
	DTC	84.78%	82.52%	89.47%	85.86%	84.62%
	RFC	84.24%	81.73%	89.47%	85.43%	84.06%
	KNN	81.52%	79.61%	86.32%	82.83%	81.36%
	GNB	85.87%	84.16%	89.47%	86.73%	85.75%

customization, making them suitable for scenarios where interpretability and fine-tuning are paramount.

Result and Discussion

AutoGluon v1.0.0 analyzes heart disease. The default settings were used as they are. AutoGluon analyzed the data to ensure it included all prediction features and had no missing values. The data was split up into 80% for training and 20% for testing by a human expert. Further, the training data was split internally into 80% and 20% of the validation subset. An AutoGluon automatically generates model training after the automated data preprocessing step. Here, the AutoGluon model training strategy is employed. This strategy used base algorithms such as KNN, RF, neural networks, and ensemble methods. Using this base algorithm, AutoGluon tested many embedded algorithms on its own and then showed the 14 best algorithms: CatBoost, RandomForestGini, RandomForestEntr, NeuralNetTorch, LightGBMLarge, NeuralNetFastAI, ExtraTreesEntr, ExtraTreesGini, LightGBM, WeightedEnsemble_L2, KNeighborsDist, XGBoost, LightGBMXT, and KNeighborsUnif. Among all these algorithms, CatBoost emerged as the best model based on its accuracy of 99.02% and ROC-AUC value of 99.89% in evaluating HF with AutoGluon. In TML, six (6) individual-trained ML models, such as LR, SVC, DTC, RFC, KNN and GNB are used. LR produced the highest accuracy in TML (87.50%), with an ROC-AUC value of 88.43%. This research compares a total of twenty (20) individual-trained ML models, consisting of fourteen (14) from AutoML and six (6) from TML, as tabulated in Table 3. Out of all 20 models, CatBoost is the best. Figure 7. depicts the performance comparison of AutoML models; similarly, figure 8. depicts the performance comparison of TML models using various performance metrics.

After evaluating the models, SHAP is employed to achieve global interpretability of the CatBoost model's predictions. Figure 9 presents the SHAP force plot of the fifth patient in the data, utilizing CatBoost. The base value (0.1122) signifies the mean SHAP value across all samples, indicating how each feature contributes to shifting the model's output from the base value. Features that elevate the prediction are highlighted in red, indicating a positive impact on the models of predicting HF, whereas those in blue diminish the prediction. Specifically, age, thal, slope, and oldpeak contribute positively to the model's HF prediction, while thalach, ca, restecg, and exang have a negative impact. The comprehensive SHAP value for each individual, denoted as f(x) = 0.53, determines whether the model predicts the presence of HF. Therefore, since f(x) is higher than the base value, the CatBoost model accurately predicts that the individual will have HF. This demonstrates the effectiveness of SHAP in enhancing model transparency and understanding, which is crucial for improving diagnostic accuracy and clinical decision-making in healthcare.

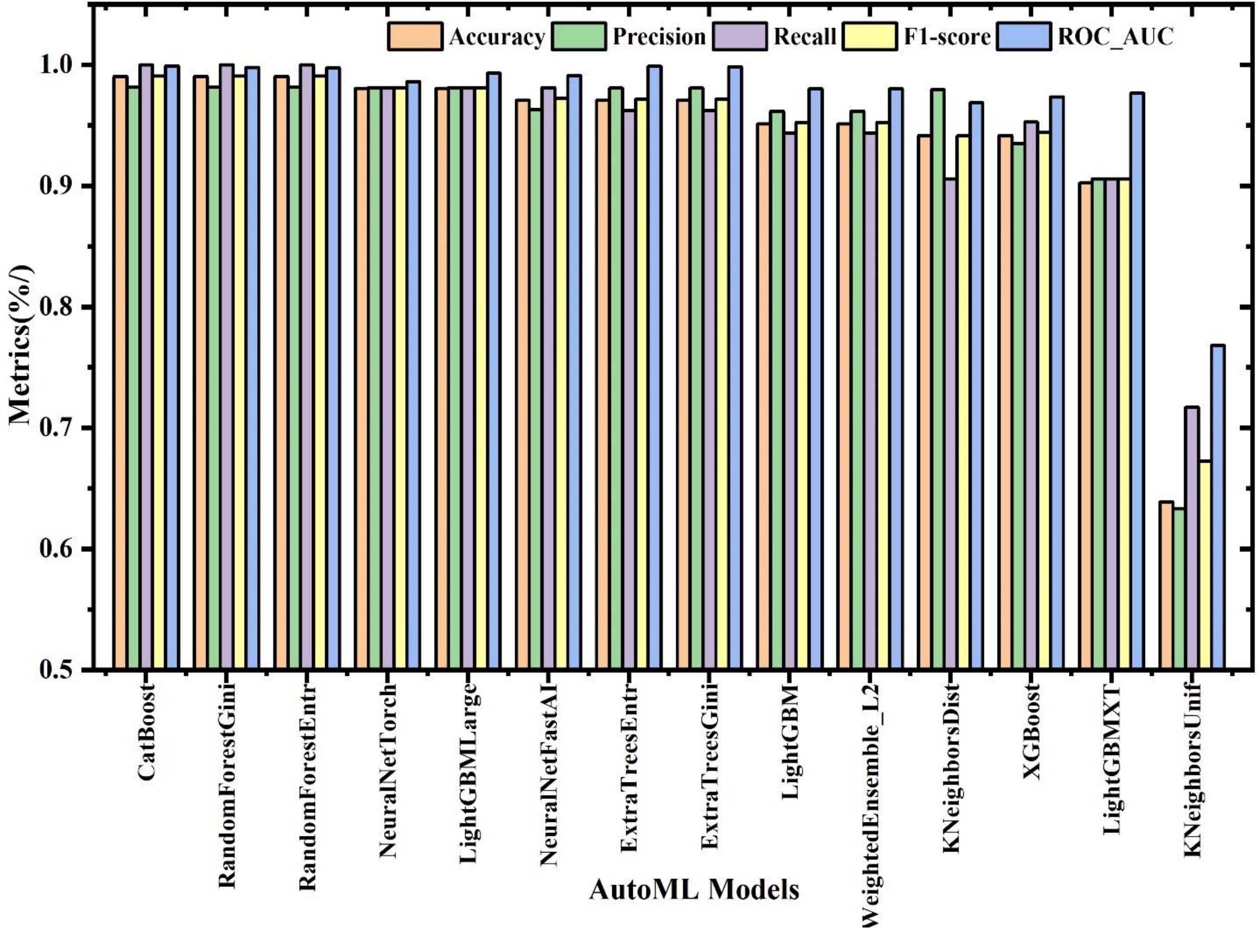


Figure 7. Performance Comparison of AutoML Models.

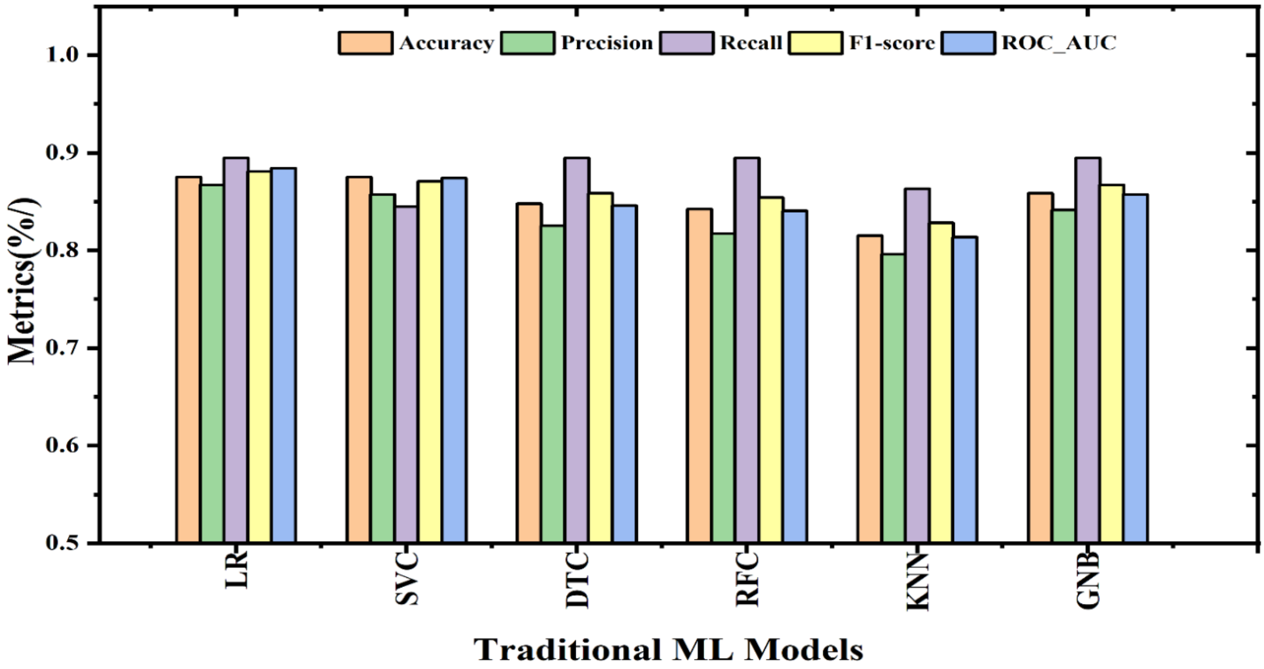


Figure 8. Performance Comparison of TML Models.

DOI: https://doi.org/10.52756/ijerr.2024.v46.003

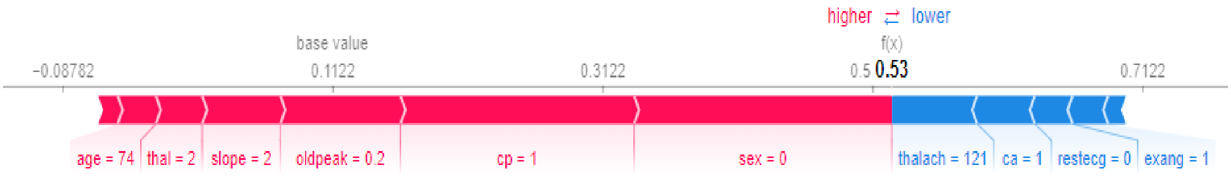


Figure 9. SHAP force plot values for 5th patient using CatBoost.

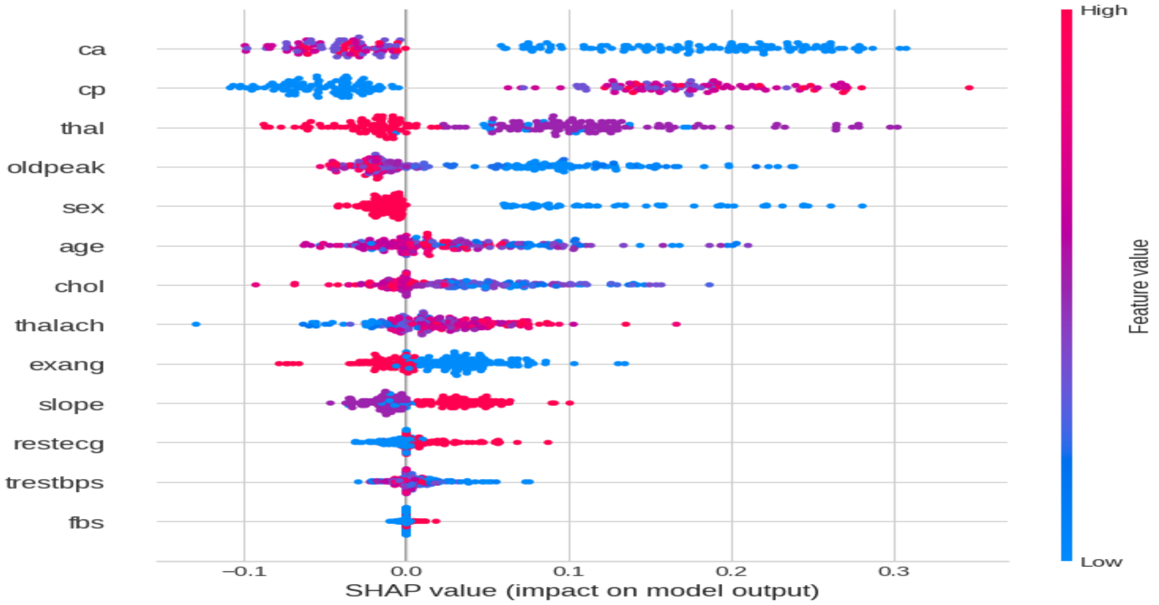


Figure 10. SHAP summary plot of the 13 attributes with the test dataset of the CatBoost.

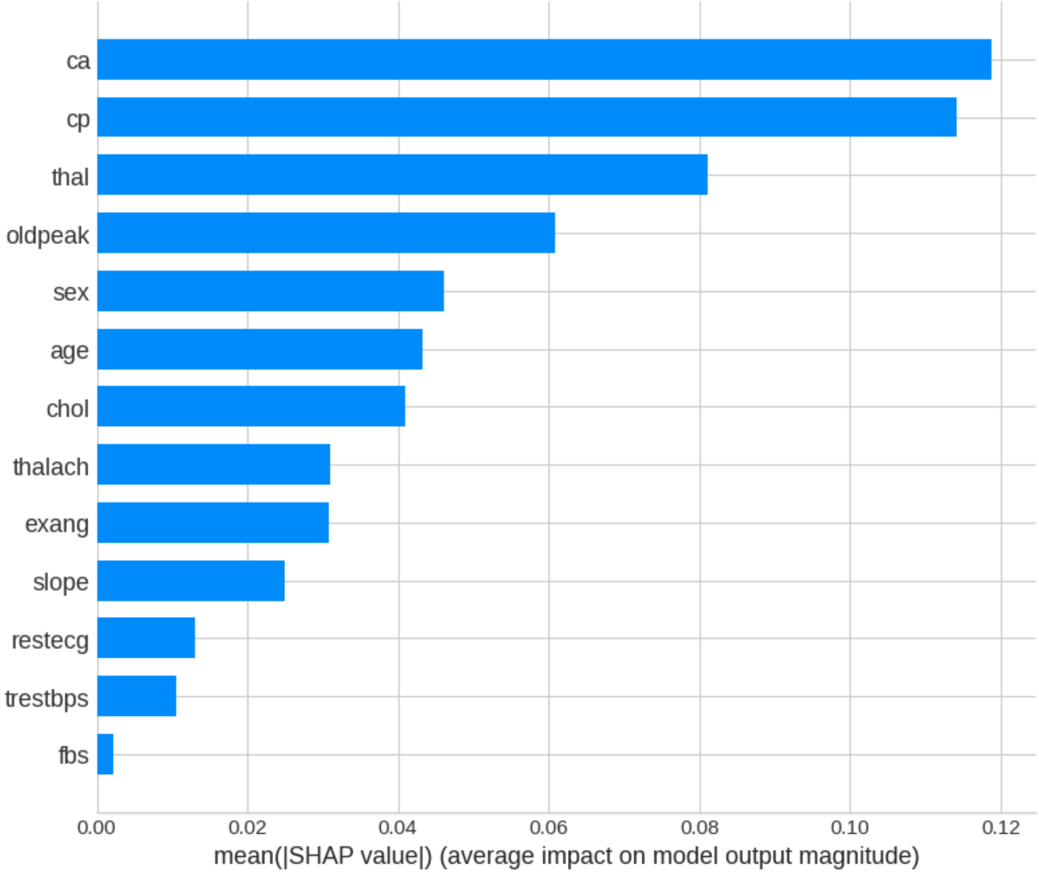


Figure 11. Features importance plot using SHAP (CatBoost).

DOI: https://doi.org/10.52756/ijerr.2024.v46.003

The CatBoost model created a summary plot of the SHAP values for 13 features from the test data, as shown in Figure 10. Each point on this plot signifies the effect of a specific feature on the model's prediction. Red points show features that positively impact the prediction when their values are high, whereas blue points indicate features that negatively impact the prediction when their values are low. The features are ordered by their average influence, showing their importance in the model's decisions. Figure 11 provides an overview of the global impact of features on HF prediction, identifying 'ca' and 'cp' as the most influential, while 'trestbps' and 'fbs' have less impact.

Conclusion

This work demonstrates the superiority of AutoML over TML models in predicting HF, highlighting the significant advantages in accuracy and efficiency. While TML models like LR achieved the highest accuracy of 87.50% and an ROC-AUC of 88.83%, they required extensive manual tuning and expert knowledge. Other TML models, such as DT, RF, KNN, and GNB, showed lower accuracy, emphasizing the optimized challenges. In contrast, the AutoML approach, particularly using the AutoGluon framework, demonstrated exceptional performance, with the CatBoost algorithm achieving an best accuracy of 99.39% and an ROC-AUC of 99.89%. The use of SHAP provided valuable interpretability, making the CatBoost model both accurate and transparent. These findings underscore AutoML's potential to streamline model selection and optimization, saving time for medical professionals and enhancing predictive accuracy. Consequently, AutoML techniques can significantly improve HF risk assessments and treatment efficiency, ultimately benefiting patient outcomes. Future research should focus on integrating AutoML frameworks with real-time clinical data, larger datasets, or multiple datasets to enhance predictive capabilities and clinical decision-making further.

Acknowledgement

I express my heartfelt gratitude to my institution for providing a conducive environment for research and study. I also sincerely thank my supervisor, Dr. Karthika Natarajan, for her invaluable guidance and support throughout this study.

Conflict of Interest

The authors declare no conflict of interest, financial or otherwise.

References

Absar, N., Das, E. K., Shoma, S. N., Khandaker, M. U., Miraz, M. H., Faruque, M. R. I., Tamam, N., Sulieman, A., & Pathan, R. K. (2022). The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction. *Healthcare*, *10*(6), 1137. https://doi.org/10.3390/healthcare10061137

Baseer, K.K., Nas, S.A., Dharani, S., Sravani, S., Yashwanth, P., & Jyothirmai, P. (2023). Medical Diagnosis of Human Heart Diseases with and without Hyperparameter tuning through Machine Learning. IEEE, In 2023 7th *International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1-8. https://doi.org/10.1109/ICCMC56507.2023.10084156

Bodapati, J., & Sajja, V. (2019). Robust Cluster-then-label (RCTL) Approach for Heart Disease Prediction. *Ingénierie Des Systèmes d Information*, *24*(3), 255–260. https://doi.org/10.18280/isi.240305

David, H., & Belcy, S. A. (2018). Heart disease prediction using data mining techniques. *Journal on Soft Computing*, *9*(1), 1824-1830. https://doi.org/10.21917/ijsc.2018.0254

Deepa, S., Prasath, S., Mohanasathiya, K. S., Ilango, M., & Ragavi, A. (2024). A Hybrid Machine Learning Approach for Enhanced Prediction of Breast Cancer with Lasso Method for Feature Extraction. *Proceedings of 4th International Conference on Artificial Intelligence and Smart Energy*, pp. 1–17. https://doi.org/10.1007/978-3-031-61471-2_1

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, *64*(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633-1650. https://doi.org/10.1111/coin.12410.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). Autogluon-tabular: Robust and accurate automl for structured data. arXiv preprint *arXiv:2003.06505.*

Ferreira, L., Pilastri, A., Martins, C. M., Pires, P. M., & Cortez, P. (2021). A comparison of AutoML tools for machine learning, deep learning and XGBoost. IEEE, In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8. https://doi.org/10.1109/IJCNN52387.2021.9534091.

Gardner, W. A. (1984). Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique. *Signal Processing*, *6*(2), 113-133. https://doi.org/10.1016/0165-1684(84)90013-6.

Gazeloğlu, C. (2020). Prediction of heart disease by classifying with feature selection and machine learning methods. *Progress in Nutrition*, 22(2).

Hajouli, S., Ludhwani, D.H.F., & Ejection Fraction. (2022). In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan. Available from: https://www.ncbi.nlm.nih.gov/books/NBK553115/.

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-based Systems*, *212*, 106622. https://doi.org/10.1016/j.knosys.2020.106622

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). Heart disease data set. *The UCI KDD Archive*. https://archive.ics.uci.edu/ml/datasets/heart+disease

Jiang, P., Suzuki, H., & Obi, T. (2023). XAI-based cross-ensemble feature ranking methodology for machine learning models. *International Journal of Information Technology*, *15*(4), 1759-1768. https://doi.org/10.1007/s41870-023-01270-2

Khourdifi, Y., & Bahaj, M. (2019). K-nearest neighbour model optimized by particle swarm optimization and ant colony optimization for heart disease classification. In Big Data and Smart Digital Environment. *Springer International Publishing*, pp. 215-224. https://doi.org/10.1007/978-3-030-12048-1_23

Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., Baber, U., ... & Narayan, S. M. (2019). Deep learning for cardiovascular medicine: a practical primer. *European Heart Journal*, *40*(25), 2058-2073. https://doi.org/10.1093/eurheartj/ehz056

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707

Nasarian, E., Abdar, M., Fahami, M. A., Alizadehsani, R., Hussain, S., Basiri, M. E., ... & Sarrafzadegan, N. (2020). Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach. *Pattern Recognition Letters*, *133*, 33-40. https://doi.org/10.1016/j.patrec.2020.02.010

National Heart, Lung and Blood Institute. (2018). Know the Differences: Cardiovascular Disease, Heart Disease, Coronary Heart Disease. Accessed August 7, 2018.

Natarajan, K., & Rajeev, C. (2024). Prediction of heart failure disease using classification algorithms along with performance parameters. In S. Kadry & S. Mahajan (Eds.), *Data Science in the Medical Field*, Academic Press, pp. 213–226. https://doi.org/10.1016/B978-0-443-24028-7.00015-5

Orlenko, A., Kofink, D., Lyytikäinen, L. P., Nikus, K., Mishra, P., Kuukasjärvi, P., ... & Moore, J. H. (2020). Model selection for metabolomics: predicting diagnosis of coronary artery disease using automated machine learning. *Bioinformatics*, *36*(6), 1772-1778. https://doi.org/10.1093/bioinformatics/btz796

Padmanabhan, M., Yuan, P., Chada, G., & Nguyen, H. V. (2019). Physician-friendly machine learning: A case study with cardiovascular disease risk prediction. *Journal of Clinical Medicine*, *8*(7), 1050. https://doi.org/10.3390/jcm8071050

Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*, *17*(1), 1100-1113. https://doi.org/10.1515/med-2022-0508

Paladino, L. M., Hughes, A., Perera, A., Topsakal, O., & Akinci, T. C. (2023). Evaluating the performance of automated machine learning (AutoML) tools for heart disease diagnosis and prediction. *AI*, 4(4), 1036-1058. https://doi.org/10.3390/ai4040053

Patil, S. B., & Kumaraswamy, Y. S. (2009). Intelligent and effective heart attack prediction system using data mining and artificial neural network. *European Journal of Scientific Research*, *31*(4), 642-656.

Pol, U. R., & Sawant, T. U. (2021). Automl: Building a classification model with PyCaret. *Ymer*, *20*, 547-552.

Purusothaman, G., & Krishnakumari, P. (2015). A survey of data mining techniques on risk prediction: Heart disease. *Indian Journal of Science and Technology*, *8*(12), 1. https://doi.org/10.17485/ijst/2015/v8i12/58385

Rajeev, C. (2024). A comparative study of Autogluonand H2O for early prediction of coronary artery disease using automated machine learning and XAI. African Journal of Biomedical Research, 5183–5193. https://doi.org/10.53555/ajbr.v27i3s.3297

Ranganathan, L. B., Rajasundaram, A., & Kumar, S. K. S. (2024). A Cross-Sectional Study on the Effect of Stress on Short-Term Heart Rate Variability and Muscle Strength Among Construction Site Workers. *International Journal of Experimental Research and Review, 44*, 1–10. https://doi.org/10.52756/ijerr.2024.v44spl.001

Rimal, Y., Paudel, S., Sharma, N., & Alsadoon, A. (2024). Machine learning model matters its accuracy: a comparative study of ensemble learning and automl using heart disease prediction. *Multimedia Tools and Applications*, *83*(12), 35025-35042. https://doi.org/10.1007/s11042-023-16380-z

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 345. https://doi.org/10.1007/s42979-020-00365-y

Shah, S. M. S., Batool, S., Khan, I., Ashraf, M. U., Abbas, S. H., & Hussain, S. A. (2017). Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. *Physica A: Statistical Mechanics and its Applications*, *482*, 796-807. https://doi.org/10.1016/j.physa.2017.04.113

Shen, Z., Zhang, Y., Wei, L., Zhao, H., & Yao, Q. (2018). Automated Machine Learning: From Principles to Practices. arXiv preprint arXiv:1810.13306. ArXiv, abs/1810.13306

Sun, B., Cui, W., Liu, G., Zhou, B., & Zhao, W. (2023). A hybrid strategy of AutoML and SHAP for automated and explainable concrete strength prediction. *Case Studies in Construction Materials*, *19*, e02405. https://doi.org/10.1016/j.cscm.2023.e02405

Tarawneh, M., & Embarak, O. (2019). Hybrid approach for heart disease prediction using data mining techniques. Springer International Publishing, In *advances in internet, data and web technologies: the 7th international conference on emerging internet, Data and Web technologies (EIDWT-2019)*, pp. 447-454.

Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. *Electronics*, *12*(8), 1789. https://doi.org/10.3390/electronics12081789

Waqar, M., Dawood, H., Dawood, H., Majeed, N., Banjar, A., & Alharbey, R. (2021). An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction. *Scientific Programming*, *2021*(1), 6621622. https://doi.org/10.1155/2021/6621622

How to cite this Article:

C. Rajeev and Karthika Natarajan (2024). Analysis and Visualize the Predictive Model Performance: Manual Vs Automated Machine Learning (AutoML) Algorithms for Heart Failure Prediction. *International Journal of Experimental Research and Review*, *44*, 31-44.

DOI : https://doi.org/10.52756/ijerr.2024.v45spl.003